Methods for Dealing with Missing Covariate Data in Epigenome-Wide Association Studies

Harriet L. Mills, Jon Heron, Caroline Relton, Matt Suderman and Kate Tilling

Correspondence address: Dr Harriet L. Mills, Medical Research Council Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Oakfield House, Oakfield Grove, Bristol, BS8 2BN, UK (email: harriet.mills@bristol.ac.uk; phone: +44 (0)117 331 0098)

Affiliations: Medical Research Council Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, UK (Harriet L Mills, Jon Heron, Caroline Relton, Matt Suderman and Kate Tilling)

Funding: This work was supported by the University of Bristol and the UK Medical Research Council (MRC, grant references MC_UU_12013/2 and MC_UU_12013/9). KT also receives funding through the MRC under grant ref: MR/M025020/1. The UK MRC and Wellcome (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for the Avon Longitudinal Study of Parents and Children (ALSPAC). ARIES (Accessible Resource for Integrated Epigenomics Studies) is funded by the Biotechnology and Biological Sciences Research Council (BBSRC, grant references BBI025751/1 and BB/I025263/1). This publication is the work of the authors (Harriet L Mills, Jon Heron, Caroline Relton, Matt Suderman, Kate Tilling) and will serve as guarantors for the contents of this paper.

Conflict of Interest: The authors declare no conflicts of interest.

Running head: Methods for Missing Covariate Data in EWAS

Abstract

Multiple imputation (MI) is a well-established method for dealing with missing data. MI is computationally intensive when imputing missing covariates with high dimensional outcome data (e.g. DNA methylation data in epigenome-wide association studies (EWAS)), because every outcome variable must be included in the imputation model to avoid biasing associations towards the null. Instead, EWAS analyses are reduced to only complete cases (CC), limiting power and potentially causing bias. We used simulations to compare five MI methods for high dimensional data under two missingness mechanisms. All imputation methods had increased power over CC analyses. Imputing separately for each variable was computationally inefficient, but dividing sites at random into evenly sized bins improved efficiency and gave low bias. Methods imputing solely using subsets of sites identified by the CC suffered from bias towards the null. However, if these subsets were added into random bins of sites the bias was reduced. The optimal methods were applied to an EWAS study with missingness in covariates. All methods identified additional sites over the CC, and many of these sites had been replicated in other studies. These methods are also applicable to other high dimensional datasets, including the rapidly-expanding area of 'omics studies.

***Keywords**: Missing data, imputation, epigenetic data, ALSPAC, ARIES*

List of Abbreviations:

MI	Multiple imputation
EWAS	Epigenome-wide Association Studies

CC	Complete Case
MM	Missingness Model
IPW	Inverse Probability Weighting

In medical research we are increasingly dealing with high-dimensional datasets detailing exposures, covariates and outcomes. This creates challenges for scaling up standard statistical methods – in terms of plausibility of underlying assumptions, but also in practicalities such as computing time. An example of this is in multiple imputation (MI) for dealing with missing data, where an approach that is practicable for a dataset with a small number of variables (fitting an imputation model 100 times for each variable and combining the results) may not be practicable for a high-dimensional dataset. Many methods for MI have been explored (1-4), but few were designed to handle datasets with over 500 covariates or where the number of covariates is much larger than the number of cases (5-7). Additionally, though MI packages (e.g. *mice* in R (8) and *ice* in Stata (9)) exist, these packages were not designed for efficiency with high dimensional datasets. Therefore, there is a need for new, efficient methods to implement MI on high dimensional datasets with missingness.

A commonly encountered high dimensional problem is that of epigenetic studies looking at DNA methylation – a reversible chemical modification of DNA whereby a methyl group is added to a cytosine nucleotide. These studies measure methylation at 480,000 or 850,000 CpG sites (per individual studied) with a few recent studies looking at millions of CpG sites, though with a small sample size (10, 11). To investigate hypotheses about the association of DNA methylation with a specific phenotype, DNA methylation measurements across the genome are tested for associations with the phenotype (often called an Epigenome-Wide Association Study, EWAS) by repeatedly fitting a simple univariate model for each CpG site – thus fitting 480,000 models in total.

Missingness in EWAS can occur in the methylation measures or the covariates. However, missingness in the methylation measures tends to be minimal and relate to technical issues of

data generation. Missingness in the covariates, by contrast, tends to have a much greater impact on the analysis, causing decreased power. Here we consider only missingness in the covariates. Commonly, a complete case (CC) analysis is used for an EWAS, where only cases with complete data on the outcome and all covariates are analysed. This will reduce the power of the analysis (in one example, the number of cases included reduced from 1018 to 678 (12)) and be biased if the chance of being a complete case is associated with the outcome, given the covariates in the model (2, 13-17). For example, if smokers are less likely to attend the clinic at which samples are taken for epigenetic analysis, and so are people with higher BMI, this would induce collider bias between smoking and BMI in a CC analysis (18). This would mean in a CC EWAS, smoking would tend to be associated with all methylation sites that were affected by BMI, and vice versa.

MI can be used to minimise bias and inefficiency in the presence of incomplete data. MI works by specifying prediction models for each variable with missingness (3). In order to avoid bias towards the null, these models need to include all variables in the analysis model (see simulation in Web Appendix 1, Web Tables 1&2 and Web Figures 1&2), plus any auxiliary variables (16, 19). For an EWAS, this would mean imputing the missing covariate data using all CpG sites, which would be computationally intensive, and require additional methods if the number of CpG sites (typically 480 000+) was greater than the number of cases (typically at most a few thousand).

Here we used simulations and an applied example to explore different methods for imputing missing values of covariates. Initially we used the computationally intensive method of imputing using each CpG site in turn. This method was extended by using groups of CpG sites together to impute the missing variable, reducing the computational time. These groups were entirely randomly selected or were designed to include some systematically selected CpGs – in this

instance using CpGs determined to be associated with the missing variable (building on work by Wu et al. (20)). We compared the imputation methods by their standard error, bias and computation time. Our conclusions are also useful for researchers analysing other high-dimensional datasets, including the rapidly-expanding area of 'omics studies.

METHODS

Simulation study

We used a publicly available dataset describing DNA methylation at 482 739 CpG sites obtained from the Human Methylation 450k array (21) for 464 individuals (22), downloaded from the Gene Expression Omnibus ((23), accession number GSE50660). Age, sex and smoking status (never, ex or current) were provided for every individual. Methylation measures were standardised for each CpG.

Covariates in this dataset had no missing data, and therefore missingness was induced for smoking status using two missingness mechanisms (MM). These were both missing at random scenarios, where missingness depended on the completely-observed covariates age and sex, so that both a CC analysis and MI including these variables as covariates would be unbiased.

MM1 – Missing with probability 75% for males aged 57 years and over.

MM2 – Missing with probability 50% for males aged 57 years and over and with probability 12.5% for all remaining individuals.

The percentages used ensured a comparable proportion of missingness for both mechanisms.

Imputation methods

All imputation models included the covariates age and sex; smoking was imputed using the "polyreg" method in mice, which uses polytomous logistic regression. 100 sets of imputed data were generated each imputation. 100 imputations is very conservative – recent literature has suggested the number of imputations should be equivalent to the percentage of missing data (a linear rule (24, 25)) or that the number is better approximated by a quadratic rule (26). Five imputation methods were used (described below and in Table 1).

Separate CpGs For each CpG in turn, smoking status was imputed using age, sex and methylation measure at that CpG site. The 100 imputed datasets for each site were pooled for the EWAS for each CpG site, using standard MI methods to obtain standard errors for the coefficients (27).

Random bins of fixed size CpG sites were divided into bins of fixed size. Smoking status was imputed for each bin using age, sex and methylation measure at all the sites in the bin. The 100 imputed datasets per bin were pooled for the EWAS analyses for CpG sites in that bin (27). Two bin sizes were used – 150 and 45 –reflecting an approximate 3:1 and 10:1 ratio of cases to variables (28, 29) and resulting in 3,219 and 10,728 bins respectively (Table 1).

The random bins method is a compromise between using every CpG in a single bin for the imputation and imputing using single CpGs in turn (the first method described above). Both are computationally intensive, with the former also having more variables than cases meaning the imputation model would not run without additional methods. Randomly assigning the CpGs into bins maximises the information being used for each imputation, while also improving the

calculation time. Other studies have also used bins to overcome the problem of many covariates ((30) being one example).

Using associated CpGs – Naïve method One multiple imputation procedure was carried out, imputing missing smoking status from age, sex and methylation measures for the set of CpG sites that were significantly associated with smoking in the CC analysis. The 100 imputed datasets were pooled for the EWAS for all CpG sites (27).

Using associated CpGs – Wu method A forward-stepwise selection model was used to select a final set of CpG sites to be included in the imputation model, from the top 100 associated CpGs identified from the CC analysis (using Bayesian Information Criterion, as Wu et al. (20)). One multiple imputation procedure was carried out, imputing missing smoking status from age, sex and methylation measures at all the selected CpG sites. The 100 imputed datasets were pooled for the EWAS for all CpG sites (27).

Random bins of fixed size always including associated CpGs – Wu bins The random binning and Wu methods were combined such that the set of CpG sites selected by the Wu method was included in every bin alongside randomly selected sites. Smoking status was imputed for each bin using age, sex and methylation measure at all sites in the bin. The 100 imputed datasets for the bin were pooled for the EWAS for CpG sites in that bin (27). As before, two bin sizes were used – 150 and 45 (including the selected sites and the random sites) resulting in 3,353 and 12,378 bins respectively (Table 1).

Ten datasets with missingness were generated for each of the two missingness mechanisms and used to perform 10 repeats of each imputation method for each mechanism. Only 10 repeats were performed because imputation and regression on such a high dimensional dataset were slow and

computationally intensive. With only 10 repeats, conclusions may be distorted by sampling variability. To confirm conclusions, the dataset was reduced to 2000 CpG sites (removing 480 739 sites) (see Web Appendix 2) and 1000 repeats were performed.

All imputation and analyses were performed on the University's High Performance Computer. Imputation and result pooling were performed using the mice and survey packages in R (31), using Rubin's rules (15, 27).

The "complete dataset" is the dataset without missingness, and the EWAS on these data gave the "truth": 298 CpG sites associated with (current and ex) smoking. The best performing method should have a high true-positive rate (the % of "true" sites correctly identified as significant by the method) and a low false-positive rate (the % of sites identified as significant by the method which were not "true" sites). Low computing time would also be advantageous. These performance measures were reported for each method, alongside the bias in the coefficients compared to the "truth".

EWAS

A linear regression analysis was used, relating age, sex and smoking status to the methylation measure at each CpG site (an EWAS):

CpG ~ age+sex+smoking

Note that this model is deliberately simplistic and does not adjust for any other covariates (such as batch effects or other confounders relevant to smoking), as the imputation methods are the focus of this simulation study.

A Bonferroni correction was used to identify those CpG sites associated with smoking (current or ex) with $P<0.05/N\_CpG$, where N_CpG was the number of CpG sites.

The EWAS was performed on the complete dataset (i.e. 464 cases with smoking, age and sex information and DNA methylation measures at 482,739 sites), to obtain a set of results representing "the truth": the 298 CpG sites associated with smoking when there was no missingness. Additionally, CC EWAS were performed for each dataset with missingness (i.e. using only those cases with complete data).

Application to an EWAS of smoking in pregnancy

The imputation methods were applied to data from the Avon Longitudinal Study of Parents and Children (ALSPAC) to illustrate their use with real missing data across multiple covariates. ALSPAC initially recruited 14 541 pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992 (32, 33), follow up increased this to 15 247 pregnancies. Detailed follow up of the mothers and children has provided a rich dataset of self-reported data, linked medical records and data collected at health clinics. The study website contains details of all the data that is available through a fully searchable data dictionary (34). Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. A sub-study, Accessible Resource for Integrated Epigenomics Studies (ARIES), selected approximately 1000 mother-child pairs and profiled DNA methylation from samples at multiple time-points in both mother and child(35).

DNA methylation was measured from blood collected at the Focus on Mothers clinic (974 cases). An EWAS explored the relationship between maternal smoking status and DNA methylation, including the following confounders believed to be associated with smoking and DNA methylation: age of the mother (at birth of the child), parity, maternal education level, housing tenure and social class; batch-effects were also included. Data taken around the birth of the child have <5% missing for individuals in ARIES, however here maternal smoking status was obtained

from a later questionnaire (around 18 years after the birth of the ALSPAC child), intentionally giving high missingness (34.6%), Web Table 3.

Three methods (random bins, Wu and Wu bins) were applied to the ARIES dataset to impute missing data for all variables. A bin size of 95 was used (10:1 ratio of cases to variables). These methods were chosen as they performed well in the simulations. Offspring birthweight, maternal alcohol intake during pregnancy and maternal smoking reported at 18weeks pregnant were also used in the imputation model but not the EWAS. The EWAS results following imputation were compared to a CC analysis, and to a review of other smoking EWAS (36).

RESULTS

Simulation study

The individuals in the simulation dataset were 70.5% male; with mean age 55.4 years (median 56); 38.6% non-smokers, 56.7% ex-smokers and the remaining 4.7% current smokers (Web Table 4)(22). The missingness mechanisms (MM1 and MM2, see Methods) gave around 22% missingness for smoking status (Web Table 4). As intended, missingness varied across age and sex (example in Web Figure 3).

For every imputation method we report the true positive (the % of "true" sites correctly identified as significant by the method) and false positive (the % of sites identified as significant by the method which were not "true" sites) rates, the mean standard errors (SEs) across betas for the CpGs (for ex-smokers) and computational time, Table 2, Figure 1 & Web Table 5. Because of the way we designed the study, *CC, separate CpGs, random bins and Wu bins* methods were unbiased (compared to the EWAS on the dataset with no missingness), Web Tables 6&7 and Web Figure 4&5. The mean SE for the Separate CpGs method is not that much smaller than that for the Complete Case method, indicating minimal improvement. However, the mean SE

decreases as more information is added to the imputation model – see for example the decrease in SE from small bins (10:1) to large bins (3:1).

The *CC* analysis had low power for both missingness mechanisms (MM1 and MM2), finding only a low number of associated CpG sites, resulting in both low true and false positive rates. The *separate CpGs* method was computationally intensive, and it correctly identified only 63.2% of the sites that were associated with smoking in the complete dataset (the true positives), with 40.5% sites false positives for MM1. The *random binning* methods were much faster than the separate CpG method. Larger bins (3:1 ratio), did not perform as well as smaller bins (10:1 ratio) which for MM1 had a high true positive rate (64.4%) and a lower false positive rate (46.1%). The *naïve* method identified a large number of associated CpG sites, and achieved the highest true positive rate of all methods (72.4%), but 65.1% of all associated sites were false positives. Associations were biased towards the null for all the "true" sites which were not identified as significant in the CC (and therefore not used in the imputation procedure), Table 3 & Web Table 8. Where the associations were strong there was less evidence of bias, Web Figures 6&7 and Web Table 9.

The *Wu* method had a relatively low false positive rate (37.3%) but did not perform quite as well as the random binning methods (61.7% true positive rate), Table 4 and Web Tables 9&10, Web Figures 6&7. The additional forward selection process meant that a much smaller number of sites (<10) were used in the imputation step. The *Wu bins* method produced very similar results to the random bins, with the smaller bins (10:1) achieving a higher true positive rate and lower false positive rate than the larger (3:1).

Results for MM2 were similar to that of MM1, though fewer associated CpG sites were found for most methods, giving correspondingly lower true and false positive rates, Table 2.

Results for each method were very similar for the 1000 repeats as for the 10 repeats on the full (unreduced) dataset and confirmed that the imputation methods involving random binning performed best, Web Table 11 & Web Figure 8. 1000 repeats showed fewer false positives in all the imputation methods (Web Table 11) (i.e. a lower rate of type 1 errors). This was particularly true in the *random binning* methods, which also identified a higher number of true positives, Web Table 11. *CC*, *separate CpGs* and the *random binning* methods were unbiased compared to the EWAS on the dataset with no missingness (Web Tables 12&13), but the *naïve* and *Wu* methods showed bias towards the null for sites not selected for the imputation model, though bias was less obvious as the reduced set of CpGs were deliberately chosen for their strong associations, Web Table 14-16 and Web Figure 9&10.

Application to an EWAS of smoking

The CC analysis identified 18 CpG sites associated with smoking in the ARIES dataset. More associations were identified when smoking status was imputed: the random binning method identified 36, the Wu method identified 29 and the Wu bins method identified 46 (Table 5 & Web Table 17). There was a large amount of overlap in the associated CpG sites identified by the four methods, Web Table 18, and 62% of all sites identified by at least one method were identified in a previous meta-analysis (36) (Table 5, Web Table 19 & Web Figure 11). This meta-analysis looked at former vs never smokers and restricted the sites to those that were differentially methylated in current vs never smokers. We note that the meta-analysis is not a gold standard, but is an indication of sites found to be related to smoking in other studies. As with the simulation study, the complete case and Wu method both tend to identify only the strongest associations – they detect fewer significant results, and those results tend to be those with the strongest associations.

DISCUSSION

Using multiple imputation to reduce the impact of missing phenotype data can improve power of EWAS studies. However, if the MI is carried out naively, bias can result. The improvement in power and detection of associated sites varied among the MI methods proposed - the optimal methods in our simulation study used random binning to reduce the number of imputations while keeping bias low. Completely random bins were simpler to implement than those including a subset of CpG sites selected using the Wu method, and performed just as well in our example. However, if some CpG sites are very strongly related to exposures or covariates, there may be benefits from including them in all imputation bins. Standard error was slightly reduced if larger bins were used, though larger bins also increased the number of falsely identified CpG sites. We have provided R code for the random bins and Wu bins methods via a Github repository (37).

The random binning method used 3:1 or 10:1 ratio of cases to variables. These ratios are generally accepted (28, 29), though the absolute limit for an imputation model is defined by the number of complete cases in the dataset. The naïve method resulted in over 150 CpG sites being selected for the imputation model: this is ~~at~~ above the upper limit of the acceptable ratio of 3:1 cases to variables in the Tsaprouni dataset and may be restrictive in other situations. Working at this upper limit may lead to overfitting, or models failing to fit. A study that has analysed optimal bin sizes for imputation, though with far fewer variables (30), found that increasing the bin size improved the imputation quality. However, there is evidence that very large bins (i.e. including many variables in the imputation model) can bias estimates towards the null when imputing an exposure (14), especially when the number of complete cases is small. Imputing for individual sites is effectively bins of size 1, and including all sites in one imputation model (not performed

here) is at the other end of the scale (1 bin of 480,000+ sites), with our bin sizes in the middle. The binning procedure is thus a careful balance between overfitting, bias and computing time.

CpG sites could be divided into bins based on their gene assignment or distance between base pairs. These methods were considered but not used (Web Appendix 2) because of the wide variety of bin sizes and the risk of collinearity. There are other variable selection methods for MI on high dimensional data (e.g. a random forest method (38)), but, like the variable selection methods evaluated here (the Wu and naïve methods) these will suffer from bias towards the null for sites that are not included in the imputation model. As the illustrative simulation showed (Web Appendix 1) where an association is strongest there will be less evidence of bias, and where an association is weakest the standard errors will be very large making it hard to distinguish bias from noise. This helps explain why the bias observed here in the Wu and naïve methods was small compared the standard errors.

Inverse probability weighting (IPW) could be used to correct the bias resulting from CC analyses, by weighting to make the set of complete cases representative (39). In theory, a high number of covariates should not be an issue for IPW, however it does rely on being able to define a model for missingness accurately. Two IPW methods were implemented (Web Appendix 2): though IPW was computationally efficient, they performed poorly with large standard errors in comparison to other methods, Web Tables 7-8&20 and Web Figures 4&5. In agreement with our results, in general, IPW is unbiased but less efficient than MI (39).

Our EWAS was not equivalent to that in Tsaprouni *et al.* (22) which adjusts for additional confounders and excludes some probes. The simulation on the Tsaprouni dataset was used to illustrate the methods and was deliberately simple with only two covariates used in the missingness mechanisms and EWAS. In reality, missingness may be a consequence of many

covariates which should all be included in the imputation model. If there were more auxiliary variables giving information on missing smoking status it is likely that imputation would be improved by including them (i.e. the imputed estimates would have smaller SEs). However, as more covariates were used (many of which had missingness), the imputation process became slower.

All imputation methods reduced the standard errors and therefore increased detection of associated CpG sites over the complete case analysis. Imputation should be used whenever missing covariate data limit the sample size for a high-dimensional dataset. Such analyses should explore sensitivity to the key assumptions: the data are missing at random, and; the imputation model has been correctly specified.

ACKNOWLEDGEMENTS

We are extremely grateful to all the families who took part in ALSPAC, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

The authors declare no conflicts of interest.

FIGURE LEGENDS

Figure 1: Scatter plots of the true positive and false positive percentages identified by the different methods for the 10 repeats on the full (unreduced) dataset. Values are listed in Table 2. Black symbols are for missingness mechanism 1 and grey are for missingness mechanism 2.

Table 1. Description and Performance of Imputation Methods.[a]

Imputation method	Summary of method		Summary of results			
	Number of imputation procedures	Imputation model	True positives	False positives	Bias	Speed (1 fastest)
Complete Case	0	NA	Poor	Very good	Unbiased[d]	1
Separate CpGs	482 739	Smoking ~ Single CpG + age + sex	Good	Good	Unbiased[d]	8
Random bins (3:1)	3219[b]	For each bin: Smoking ~ 150 CpGs + age + sex	Good	Poor	Unbiased[d]	6
Random bins (10:1)	10 728[b]	For each bin: Smoking ~ 45 CpGs + age + sex	Good	Good	Unbiased[d]	4
Naive Method	1	Smoking ~ CC CpGs + age + sex	Good	Poor	Biased towards the null for non-CC CpGs	3
Wu Method	1	Smoking ~ Selected CpGs + age + sex	Good	Good	Biased towards the null for non-selected CpGs	2
Wu bins (3:1)	3353[b,c]	For each bin: Smoking ~ 150 CpGs (including Wu selected CpGs) + age + sex	Good	Poor	Unbiased[d]	7
Wu bins (10:1)	12 378[b,c]	For each bin: Smoking ~ 45 CpGs (including Wu selected CpGs) + age + sex	Good	Good	Unbiased[d]	5

Abbreviations: CC: Complete Case

[a]This table summarises the imputation methods described in the text and their results for the simulations only. N_{CpG} is the number of CpG sites included in the analysis (482,739).

[b]This is approximately (N_{CpG} / binsize).

[c]Recall that the bins for the Wu bins method always contain the subset of CpGs selected in the forward stepwise process, so there are slightly more bins for the Wu bins method than the random bins method to accommodate the extra sites.

[d]These methods are only unbiased if the imputation model is correct and data are MAR

Table 2.: Detailed Comparison of Imputation Methods Performance for the 10 repeats.[a]

Imputation method	MM1						MM2						Time relative to separate CpG method
	Associated CpGs					SE of coefficients (mean (SD))	Associated CpGs					SE of coefficients (mean (SD))	
	Total	# in complete	% of complete	# not in complete	% of total found		Total	# in complete	% of complete	# not in complete	% of total found		
Complete Data	298					0.0953 (0.0081)	298					0.0953 (0.0081)	
Complete Case	169.7	139.3	46.7	30.4	16.7	0.1073 (0.0099)	147.1	122.5	41.1	24.6	13.6	0.1069 (0.0099)	0.002
Separate CpGs	330	188.3	63.2	141.7	40.5	0.1052 (0.0093)	282.8	169.9	57.0	112.9	34.1	0.1049 (0.0093)	1.000
Random bins (3:1)	537.2	189.6	63.6	347.6	63.5	0.0997 (0.0086)	482.2	170.5	57.2	311.7	58.8	0.0985 (0.0085)	0.517
Random bins (10:1)	373.6	192	64.4	181.6	46.1	0.1031 (0.0090)	326.3	176.6	59.3	149.7	38.9	0.1028 (0.0090)	0.339
Naive Method	863.3	215.8	72.4	647.5	65.1	0.0984 (0.0084)	433.9	180.1	60.4	253.8	45.2	0.0974 (0.0083)	0.069
Wu Method	312	183.8	61.7	128.2	37.3	0.1002 (0.0087)	290.4	170.8	57.3	119.6	28.2	0.1001 (0.0087)	0.059
Wu bins (3:1)	516.7	196.9	66.1	319.8	59.9	0.0984 (0.0084)	432.6	175.2	58.8	257.4	53.4	0.0972 (0.0083)	0.527
Wu bins (10:1)	410.2	202.2	67.9	208	48.2	0.0996 (0.0086)	349.2	187.1	62.8	162.1	38.3	0.0995 (0.0086)	0.412

[a] The number of CpG sites associated with smoking (ex or current) identified as significant in the regression analysis for each method, for both missingness mechanisms (MM). We report the number of these which were also significant in the EWAS on the complete dataset (presented with the percentage, i.e. the true positive rate) and the number of those which were not significant in the EWAS on the complete dataset (presented with the percentage of those found to be significant which were "incorrect", i.e. the false positive rate). The mean and SD of the standard errors (for the coefficients for the association of each CpG with being an ex-smoker) are reported for each method. Note that this is the mean across repeats of the mean and SD of the standard errors (SE) within each repeat. Recall that the analysis on the complete data and CC analysis did not require any imputation, making their computation time very low. Relative times are calculated from computation times averaged over example runs for MM1 and MM2, raw times are provided in the Web Table 5. The table shows the results of the 10 repeats on the full (unreduced) dataset.

Table 3. Naïve Method Performance Details for Ex-Smokers[a].

Scenario	CpG sites identified as significant in CC				CpG sites identified as significant in analysis on the complete data and not in CC analysis				All other CpG sites			
	Positive		Negative		Positive		Negative		Positive		Negative	
	Mean (SE)	Mean bias (SD)	Mean (SE)	Mean bias (SD)	Mean (SE)	Mean bias (SD)	Mean (SE)	Mean bias (SD)	Mean (SE)	Mean bias (SD)	Mean (SE)	Mean bias (SD)
MM1												
N	63.2		106.5		85.0		73.7		235482.8		246927.8	
Complete Data ("truth")	0.5641 (0.09207)		-0.5962 (0.09115)		0.5261 (0.09249)		-0.5239 (0.09252)		0.1479 (0.09473)		-0.1468 (0.09578)	
Wu Method	0.6187 (0.09473)	0.0546 (0.0541)	-0.6400 (0.09405)	-0.0438 (0.0603)	0.5187 (0.09602)	-0.0074 (0.0505)	-0.5150 (0.09609)	0.0089 (0.0530)	0.1625 (0.09782)	0.0147 (0.0568)	-0.1611 (0.09892)	-0.0143 (0.0559)
MM2												
N	54.2		92.9		91.7		83.8		235485.1		246931.3	
Complete Data ("truth")	0.5562 (0.09214)		-0.6010 (0.09091)		0.5301 (0.09242)		-0.5287 (0.09252)		0.1479 (0.09473)		-0.1468 (0.09578)	
Wu Method	0.5722 (0.09424)	0.0160 (0.0530)	-0.6136 (0.09338)	-0.0126 (0.0586)	0.4983 (0.09517)	-0.0318 (0.0479)	-0.4978 (0.09529)	0.0309 (0.0498)	0.1465 (0.09686)	-0.0013 (0.0481)	-0.1466 (0.09794)	0.0002 (0.0485)

Abbreviations: CC: Complete Case

[a]Average beta (average standard error in brackets) and average bias (standard deviation in brackets) for ex-smokers specifically for the Naive method compared to the EWAS on the complete dataset (N=263). The table shows the results of 10 repeats on the full (unreduced) dataset. CpG sites have been divided into three groups: (1) CpG sites identified as significant in the CC, (2) CpG sites identified as significant in the complete dataset and not in the CC and (3) all other CpG sites. We divide the beta into positive (>0) and negative (<0) according to their value in the EWAS on the complete dataset. Web Table 8 is the equivalent for current smokers (N=22).

Table 4. Wu Method Performance Details for Ex-Smokers[a].

Scenario	CpG sites selected from CC				CpG sites identified as significant in analysis on the complete data and not selected				All other CpG sites			
	Positive		Negative		Positive		Negative		Positive		Negative	
	Mean (SE)	Mean bias (SD)	Mean (SE)	Mean bias (SD)	Mean (SE)	Mean bias (SD)	Mean (SE)	Mean bias (SD)	Mean (SE)	Mean bias (SD)	Mean (SE)	Mean bias (SD)
MM1												
N[b]	1.4		6.2		134.9		156.5		235494.7		246945.3	
Complete Data ("truth")	0.6112 (0.0907)		-0.7925 (0.0857)		0.5465 (0.09223)		-0.5666 (0.09173)		0.1479 (0.09473)		-0.1469 (0.09578)	
Wu Method	0.6616 (0.09551)	0.0504 (0.0272)	-0.8456 (0.08873)	-0.0531 (0.0357)	0.5411 (0.09744)	-0.0054 (0.0441)	-0.5743 (0.09622)	-0.0077 (0.0408)	0.1606 (0.09961)	0.0127 (0.0471)	-0.1591 (0.1008)	-0.0122 (0.0462)
MM2												
N[b]	1.8		6.2		134.4		156.3		235494.8		246945.5	
Complete Data ("truth")	0.5735 (0.09118)		-0.7211 (0.08725)		0.5465 (0.09222)		-0.5690 (0.09166)		0.1479 (0.09473)		-0.1469 (0.09578)	
Wu Method	0.6313 (0.09471)	0.0578 (0.0513)	-0.7522 (0.09058)	-0.0311 (0.0424)	0.5362 (0.09736)	-0.0103 (0.0482)	-0.5648 (0.09632)	0.0042 (0.0423)	0.1559 (0.09948)	0.0080 (0.0461)	-0.1552 (0.1007)	-0.0083 (0.0461)

Abbreviations: CC: Complete Case

[a]Average beta (average standard error in brackets) and average bias (standard deviation in brackets) for ex- smokers specifically for the Wu method compared to the EWAS on the complete dataset (N=263). The table shows the results of 10 repeats on the full (unreduced) dataset. CpG sites have been divided into three groups: (1) CpG sites selected by Bayesian Inference Criterion (BIC) from those identified as significant by the CC, (2) CpG sites identified as significant in the EWAS on the complete dataset which were not selected by the BIC and (3) all other CpG sites. We divide the beta into positive (>0) and negative (<0) according to their value in the EWAS on the complete dataset. Web Table 10 is the equivalent for current smokers (N=22).

[b]Average number of CpG sites in that group, across the 10 repeats.

Table 5. Associations in ARIES by Imputation Method.[a]

Imputation method	Number of CpG sites identified by the imputation method	% of the sites identified by the imputation method that were also reported in Joehanes		% of the 185 sites identified after BC in Joehanes, that were identified by the imputation method
		In the 2568 sites	In the 185 sites after BC	
Complete Case	18	94.4	83.3	8.1
Random Bins	36	72.2	50.0	9.7
Wu Method	29	93.1	82.8	13.0
Wu Bins	46	63.0	47.8	11.9
Total unique	60	61.7	43.3	14.0

Abbreviations: ARIES: Accessible Resource for Integrated Epigenomics Studies, BC: Bonferroni Correction

[a] Number of CpGs identified as significantly associated with smoking (ex- or current) in the ARIES dataset, using the complete case analysis and three imputation methods. Reported are the number of CpGs which were significantly associated, the percentage of these which were replicated in the 2,568 CpG sites reported by Joehanes et al (current versus never smokers (FDR<0.05)), the percentage of those which were replicated in the 185 CpGs reported by Joehanes et al after Bonferroni Correction, and the percentage of those 185 which were identified by each method.

REFERENCES

1. Carpenter J, Kenward M. *Multiple imputation and its application*. John Wiley & Sons, Hoboken, New Jersey; 2012.
2. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393.
3. Van Buuren S. *Flexible imputation of missing data*. CRC Press, Boca Raton, Florida; 2012.
4. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine* 2010;29(28):2920-31.
5. Deng Y, Chang C, Ido MS, et al. Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data. *Scientific Reports* 2016;6:21689.
6. Liao SG, Lin Y, Kang DD, et al. Missing value imputation in high-dimensional phenomic data: imputable or not, and how? *BMC Bioinformatics* 2014;15(1):1-12.
7. Zhao Y, Long Q. Multiple imputation in the presence of high-dimensional data. *Stat Methods Med Res* 2016;25(5):2021-35.
8. van Buuren SaG-O, Karin. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 2011;45(3):1-67.
9. Royston P. ICE: Stata module for multiple imputation of missing values. *Statistical Software Components*: Boston College Department of Economics, 2006.
10. Klughammer J, Kiesel B, Roetzer T, et al. The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. *Nature Medicine* 2018;24(10):1611-24.
11. Rizzardi LF, Hickey PF, DiBlasi VR, et al. Neuronal brain-region-specific DNA methylation and chromatin accessibility are associated with neuropsychiatric trait heritability. *Nature Neuroscience* 2019;22:307-16.
12. Küpers LK, Xu X, Jankipersadsing SA, et al. DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. *International Journal of Epidemiology* 2015;44(4):1224-37.
13. Bartlett JW, Frost C, Carpenter JR. Multiple imputation models should incorporate the outcome in the model of interest. *Brain* 2011;134(Pt 11):e189.
14. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods* 2001;6(4):330-51.
15. Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Stat Methods Med Res* 2007;16(3):199-218.
16. Moons KG, Donders RA, Stijnen T, et al. Using the outcome for imputation of missing predictor values was preferred. *Journal of clinical epidemiology* 2006;59(10):1092-101.
17. Spratt M, Carpenter J, Sterne JA, et al. Strategies for multiple imputation in longitudinal studies. *Am J Epidemiol* 2010;172(4):478-87.
18. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;15(5):615-25.
19. Little RJ. Regression with missing X's: a review. *Journal of the American Statistical Association* 1992;87(420):1227-37.
20. Wu C, Demerath EW, Pankow JS, et al. Imputation of missing covariate values in epigenome-wide analysis of DNA methylation data. *Epigenetics* 2016;11(2):132-9.
21. Infinium HumanMethylation450 BeadChip Kit. (http://www.illumina.com/products/methylation_450_beadchip_kits.html). (Accessed October 26 2016).

22. Tsaprouni LG, Yang TP, Bell J, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics* 2014;9(10):1382-96.
23. NCBI Gene Expression Omnibus. (http://www.ncbi.nlm.nih.gov/geo/). (Accessed July 21 2016).
24. Bodner TE. What improves with increased missing data imputations? *Structural Equation Modeling* 2008;15(4):651-75.
25. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 2011;30(4):377-99.
26. von Hippel PT. How Many Imputations Do You Need? A Two-stage Calculation Using a Quadratic Rule. *Sociological Methods & Research* 2018:0049124117747303.
27. Little RJ, Rubin DB. *Statistical analysis with missing data*. John Wiley & Sons, Hoboken, New Jersey; 2014.
28. Hardt J, Herke M, Leonhart R. Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC medical research methodology* 2012;12(1):1.
29. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 1996;49(12):1373-9.
30. Yin X, Levy D, Willinger C, et al. Multiple imputation and analysis for high-dimensional incomplete proteomics data. *Statistics in Medicine* 2016;35(8):1315-26.
31. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2016.
32. Boyd A, Golding J, Macleod J, et al. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* 2013;42(1):111-27.
33. Fraser A, Macdonald-Wallis C, Tilling K, et al. Cohort profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International Journal of Epidemiology* 2012;42(1):97-110.
34. Avon Longitudinal Study of Parents and Children: Access data and samples. (http://www.bristol.ac.uk/alspac/researchers/access/). (Accessed July 31 2019).
35. Relton CL, Gaunt T, McArdle W, et al. Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *Int J Epidemiol* 2015;44(4):1181-90.
36. Joehanes R, Just AC, Marioni RE, et al. Epigenetic Signatures of Cigarette Smoking. *Circulation: Cardiovascular Genetics* 2016;9(5):436-47.
37. Code released with the paper "Methods for dealing with missing covariate data in EWAS studies". (https://github.com/harrietlmills/MethodsMissingCovariateData). (Accessed July 31 2019).
38. Shah AD, Bartlett JW, Carpenter J, et al. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American Journal of Epidemiology* 2014;179(6):764-74.
39. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res* 2013;22(3):278-95.